综合型科学数据仓储元数据创建服务研究

■ 黄国彬 王涛
北京师范大学政府管理学院 北京 100875

摘 要:[目的/意义]缺乏科学数据元数据知识与高效易用的元数据创建服务阻碍了科研人员科学数据的共享与重用,综合型科学数据仓储因数据存储量大、面向用户广,其所提供的元数据创建服务对改善上述困境具有借鉴意义。[方法/过程]以Springer Nature和Scientific Data推荐的6个综合型科学数据仓储为样本,从服务的内容构成与服务的实现模式两个方面对其元数据创建服务进行调研分析,归纳其服务特点与先进经验。[结果/结论]综合型科学数据仓储所提供的元数据创建服务具有沿袭传统并有所创新、力求简洁并凸显自身特色、重视元数据知识普及与能力转化、充分保证数据民主、注重关联资源组织并鼓励数据引用五大特点,其服务模式既重视服务的易用性、有用性又注重元数据知识普及,对我国图情机构数据仓储建设及元数据创建服务开展具有重要启示和借鉴意义。

关键词:科学数据 元数据创建服务 综合型仓储
分类号:G239.2
DOI:10.13266/j.issn.0252-3116.2021.21.020

1 引言

高效的科学数据管理需要强有力的技术支持但更依赖于高质量的数据组织,高质量的数据组织取决于描述数据的元数据的完备性和系统性。数据集发布是存储数据集及其相关属性以使其可供用户社区访问的过程。作为数据集发布的起点,用户借助平台内元数据创建服务主动提交其发布所需的元数据,其服务水平与质量不仅影响科研人员发布并共享其数据集的意愿,而且直接决定所发布数据集的元数据质量。元数据作为数据仓储所提供的数据浏览、检索、共享等服务的基础,如何为用户提供优质的元数据创建服务是科学数据仓储建设过程中需要考虑的核心问题之一。

2 国内外研究综述

为深入了解科研人员共享科学数据的动机与方式,Wiley[1]的一项研究在调查了全球2 886名科研人员后总结出科研人员不愿共享其科学数据的4个重要原因:①害怕共享数据产生的诸如数据滥用、法律或者商业后果等负面影响;②缺乏对其工作的认可机制;③发布数据所涉及的准备工作量太大;④缺少关于如何以及在何处共享数据的知识。T. CAROL[2-3]和其团队在2011年和2015年分别对1 000多名科研人员如何管理其科研数据进行调查,在2011年的调查结果中超过50%的科研人员表示从未使用过任何元数据标准且仅26%的科研人员对自己所使用的科学数据元数据创建工具感到满意,在2015年的调查结果中仍然有47.9%的研究人员表示从未使用过任何元数据标准。《开放数据状态报告2019》[4]表明全球仍有54.33%的科研人员从未听说过FAIR(Findability, Accessibility, Interoperability, and Reusability)原则,有48%的科研人员不清楚如何运用数据许可协议。国外相关研究表明:当前科研人员缺少科学数据元数据知识且不擅于为自己的科学数据创建元数据,元数据创建服务与工具存在短板,上述原因直接或间接影响了科学数据的共享与重用。

在CNKI中使用标题检索,构造检索式“(篇名:(元数据 + 科学数据))AND(篇名:服务)”,最终得到89篇相关文献,采用由上到下逐层缩小范围的方法对文献进行梳理发现国内研究呈现如下特点:①较多关注国外高校图书馆内所开展科学数据服务情况,侧重于整体分析并引进国外先进经验,如肖潇等[5]归纳出

作者简介:黄国彬(ORCID:0000-0001-9059-8285),副教授,博士,硕士生导师,;王涛(ORCID:0000-0002-4398-2281),硕士研究生,通讯作者,E-mail:13051817211@163.com。
收稿日期:2021-06-06 修回日期:2021-09-11 本文起止页码:131-140 本文责任编辑:易飞

在 E-science 环境下国外图书馆参与科学数据服务的 5 种服务方式,王翠萍等[6]以美国、英国和澳大利亚 5 所高校为例总结出高校图书馆 7 大科学数据服务项目;②较早关注科学数据在描述和组织方面可能面临的分类、标引、文献关联等问题,相关学者调研并总结了国外利益相关方探索与解决上述问题的技术措施与应对方案,如钱鹏等[7]提出科学数据组织与服务需解决的六大关键问题,邱春燕[8]总结了国外期刊文献与科学数据关联服务的提供途径与关键性实现方式;③少数学者关注了国外高校图书馆提供的元数据服务及元数据创建服务,如黄鑫等[9]调查分析了 8 所 USA News 排名前 100 位的高校图书馆科学数据元数据服务并总结出其元数据创建服务的 4 种形式;④部分学者对典型数据仓储的元数据方案进行了研究,未对平台内如何将元数据标准或方案运用于具体的元数据创建服务中进行研究,如胡芳[10]研究了国外 4 个典型数据仓储的元数据方案。

总体来看,国内有关科学数据元数据创建服务的研究较少,相关研究主要从宏观层面介绍、分析并引进国外高校图书馆科学数据服务的先进模式与经验,对元数据创建服务的少数研究也局限在高校图书馆内,缺乏对外部平台元数据创建服务的系统分析与经验总结。本文选取 6 个综合型科学数据仓储,聚焦平台所提供的元数据创建服务,结合相关学者所提出的科学数据组织、文献关联等问题,分析并总结其元数据创建服务的特点,以期为我国高校图书馆数据仓储建设及元数据创建服务的设计与实施提供启示与借鉴,最终促使其为科研人员提供易用且高质量的科学数据元数据标注方案。

3 调研对象与研究路线

科学数据仓储(research data repository,简称 RDR)又被称为数据仓储、中心、平台等,作为科学数据管理的基础设施,一般分为学科型和综合型。根据 Springer Nature 旗下学术期刊对数据提交的要求[11],学科型仓储是论文关联数据根据其论文所属的特定学科提交至学界认可的仓储,综合型仓储是当数据无合适的学科仓储供提交时所用的备选仓储。在学科交叉融合背景下,相比于学科型数据仓储,综合型数据仓储服务群体更为广泛,一般面向整个科学共同体提供科学数据的创建、提交、存储、出版和管理服务,其支持任何类型、任何学科内科学数据的存储,因而建设的要求和标准更高,能充分体现一国的科学数据共享基础设施的建设与服务水平。

本研究选取 Springer Nature 与 Scientific Data 推荐的综合型科学数据仓储名单上的 6 个仓储作为调查对象[12]:Dryad Digital Repository(以下简称 DDR)、Figshare、Harvard Dataverse(以下简称 HD)、Zenodo、Mendeley Data(以下简称 MD)、Science Data Bank(以下简称 SDB),平台基本信息如表 1 所示(数据收集日期截止到 2021 年 1 月)。选取这些仓储的原因在于:①平台内元数据创建服务与其他服务边界清晰且具有独立性;②平台运营主体具有多样性,不局限于高校图书馆;②平台建设成熟且面向全球科研人员提供服务,收录数据量较大,表明平台服务质量和能力受到学界认可,其元数据创建服务具有参考价值;④2020 年 9 月 SDB 成为该名单上唯一一个中国自主研发建设的仓储[13],所选仓储同时具有国内、国际代表性。采用网络调研并结合文献调研,本文将从服务的内容构成、服务的实现模式两个方面剖析 6 个平台的元数据创建服务。

表 1 数据仓储基本信息

数据仓储	国家	运营主体	已收录数据/条
DDR	美国	CDL	38 828
Figshare	英国	Figshare LLP	5 668 636
HD	美国	哈佛大学	1 015 374
Zenodo	欧盟	CERN	1 714 038
MD	荷兰	Elsevier	27 090 178
SDB	中国	中国科学院	1 055

注:CDL 指 California Digital Library(加州大学数字图书馆),CERN 指 the European Organization for Nuclear Research(欧洲核子研究组织)

4 元数据创建服务的内容构成

4.1 元数据元素及其值的设置

平台元数据元素设置(见表 2)主要与数据仓储建设之初所参照元数据框架有关,数据仓储在制定元数据方案时通常会参考一个或者多个元数据标准形成平台自身的元数据方案并根据需要进行元素及其值的扩展或缩减。Zenodo 和 DDR 明确说明其元数据元素设置基于 DataCite schema,HD 在元数据元素设置上主要采纳和借鉴了 DDI(Data Documentation Initiative)、DC(Dublin Core)、DataCite schema 这三个元数据框架,其余 3 个平台的官网中虽未明确说明其元数据元素设置所依据的元数据框架,但是其具体的元数据元素设置表明 3 者在力求简洁的基础上充分吸收和借鉴了自然科学[14]和社会科学[15]领域相关元数据标准。

表 2　平台元数据元素

数据仓储	必备元素	选择性元素
DDR	刊名及文章 ID、标题、作者、摘要	关键词、方法、使用说明、关联作品、研究领域、资助
Figshare	标题、作者、分类、项目类型、关键词、描述、许可协议	资助、参考、数据保护(＊配合许可协议使用,包括时间设定、保护范围、原因说明三个元素)
HD	数据集层:标题、作者、联系方式、描述、主题 文件层:无	数据集层:关键词、关联出版物、主题分类、说明、语言、关联材料、关联数据集、存储日期、存储者、访问限制、时间范围、数据搜集、分发日期等 文件层:名称、路径、描述、标签、溯源文件、访问限制等
Zenodo	出版日期、标题、作者、描述、访问权限、许可协议、关键词	语言、附加说明、资助、关联出版物与数据集、贡献者、参考文献、主题
MD	标题、贡献者、描述、许可协议	机构、重现步骤、参考及关联资源、保护期设定
SDB	语言、标题、关键词、数据集简介、学科分类、通讯作者邮箱、数据集作者、数据类型、许可协议、数据共享方式	关联论文标题、URL、DOI,数据集参考链接、基金支持信息

6 个平台元数据元素设置模式基本相同(见图 1),其元素总体上可分为必备元素和选择性元素,其中必备元素是数据集提交与发布所必需的元素,选择性元素是满足数据发布要求之外用户根据自身需要为数据集附加的相关属性。必备元素和选择性元素又可进一步分为复合元素与单一元素,其中复合元素是指包含至少两个子元素的元素,如作者这一必备元素通常包含姓名、所属单位、ORCID、邮箱等子元素,而关联作品这一选择性元素通常包含关联资源标识符或 URL 链接、关系、资源类型这三个子元素。单一元素与复合元素又可分为可重复与不可重复两个大类,可重复元素是指该元素可重复著录多次,6 个平台中作者这一元素均为可重复复合元素。在复合元素下,无论可重复与否,其子元素同样可分为必备子元素与选择性子元素。

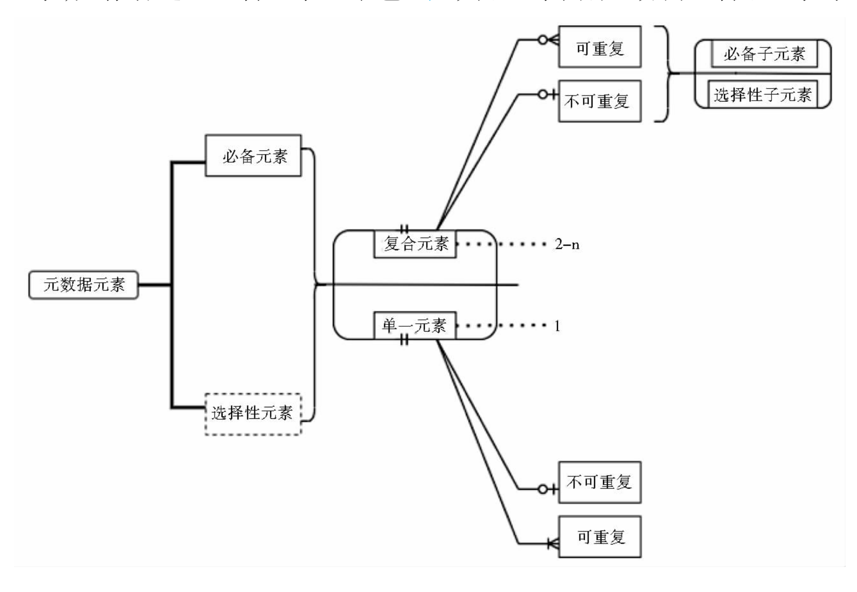


图 1　元数据元素设定模式

值的设置既需使得数据集的描述具有完备性又要考虑用户的易用性,还需考虑值的规范控制。6 个平台数据元素值可分为三类:第一类为固定的选择性值,以下拉列表形式供用户直接选择;第二类为外部参考值,提供外部参考标准供用户选择著录;第三类为用户著录值,无任何参考信息或者仅给出著录要求,由用户根据已有知识或者阅读元素解释、指导文档后直接著录。前两类值的设定,参考后控词表或者领域本体,并从不同术语系统中吸收元数据属性值,为数据集的描述提供了规范参考和有形约束,第三类值则充分体现了开放、自由原则。

4.2　元数据元素组织方式

元数据元素组织是将原本松散无序的元数据元素、元素值根据功能和属性的不同进行重新组合与排序的过程。作为将用户为数据集创建的初始元数据转换为平台最终生成的数据集标准元数据这一过程的枢纽,其具有两大功能:①用户指导功能,使得用户可以快速理解平台内科学数据元数据创建与整合的逻辑框架;②中介转换功能,以元数据元素组织框架形成用户元数据创建模板,以该模板为中介最终将用户为数据集创建的初始元数据转换为平台所发布的数据集标准元数据。

笔者在充分调研了自然科学和社会科学领域典型数据集元数据标注与组织方案并选取领域标准案例进行对比（结果见表 3）后，发现二者虽因学科领域不同造成标注与组织的侧重点不同，但其元数据元素设置与组织的整体框架基本相同且呈现标准化趋势，即通过标准的模块化元数据组织方式实现元数据元素的系统整合与资源的内外联动。

表 3　社会科学与自然科学领域数据集元数据标注示例

数据集信息	劳动力调查[16]	1978 – 2016 年中国种植业温室气体排放量[17]
来源	UK Data Service	Scientific Data
学科领域	社会科学	自然科学
标注细节	模块 1：细节 子模块：必要细节（标题、研究编号、获取协议、DOI、系列、主要调查机构）、资助者与贡献者（数据提交者、资助者）、引用和版权（引用格式选择、XML 引用格式、版权说明）、主题、主题词表下的关键词、摘要（背景、内部文件说明、数据加权声明、衍生数据说明、终端用户许可协议与安全访问说明、内部变量补充、去除、变化的说明、调查方法说明、变量加权信息、问卷设计说明）、覆盖范围和方法（调查的起始日期、国家、空间单位、调查个体、调查的地理范围、人口、样本数量、数据收集方法、时间维度、抽样程序、加权方法）、版本历史（初版发布日期、最新版本发布日期、所有版本）	模块 1：单元 子模块：摘要（数据集整体概述、测量对象、技术类型、因素类型、样本特征 – 环境、样本特征 – 地点）、背景与总结、方法（作物露天秸秆露天焚烧、水稻种植、耕地变化、耕地排放、农药生产等）、数据记录、技术验证、局限性、代码可获得性、参考资源、作者信息（附属机构、通讯作者）、道德声明（竞争利益）、附加信息、补充资料、版权与许可、关于本数据集（数据集引用、数据集接收日期、接受日期、发布日期、DOI、分享链接、主题词）
	模块 2：内部文件（标题、文件名称、大小）	模块 2：数据集内部图表（图表标题、简短的解释与说明）
	模块 3：参考资源（案例研究、出版物/报告、调查研究、其他）	模块 3：参考资源（＊统一采用标准引用格式）

综合型数据仓储元数据组织方式并没有完全照搬自然科学与社会科学领域标准的模块化元数据元素组织方式，而是根据自身元数据元素数量、复杂程度与平台功能定位进行调整、取舍并有所创新（见表 4）。平台主要采用三种元数据组织方式：①元素数量较少的采用直接罗列方式，意在表现其简洁性；②元素数量中等的则采用模块化组织，意在体现其组织的逻辑性；③元素数量较多且复杂的平台如 HD 则采用创新的外部分层与内部模块化组织方式，意在将元素组织的逻辑性与层次性结合，并完成数据集从宏观到微观层次的充分描述以实现从数据集整体到局部数据文件的引用。无论采取何种组织方式，平台的最终目的都是从用户的易用性出发，使其为科学数据创建元数据时能够沿着一条清晰主线，高效且规范地完成元数据著录。

表 4　平台元数据元素组织方式

数据仓储	元素数量	组织方式	说明	组织方式特点
Figshare	较少	直接罗列	——	简洁明了
MD	较少	直接罗列	——	
DDR	中等	模块化组织	四大模块：初步信息、数据集基本信息、数据描述、关联作品	
Zenodo	中等	模块化组织	三大必备模块：数据集类型、基本信息、许可协议；五大选择性模块：资助、关联/替代标志符、贡献者、参考资源、关联资源	注重整体逻辑性、强调模块内部元素功能一致性、保持模块之间关联性
SDB	中等	模块化组织	五大模块：数据集描述信息、数据集作者信息、论文关联信息、数据权益相关信息、数据实体文件	
HD	较多且复杂	外部分层组织＋内部模块化组织	将数据集描述分为宏观数据集层描述与微观文件层描述，在数据集层内部进行元素的模块化组织	凸显每一个元数据元素在描述功能的层次适用性，实现数据集的充分描述并融合元素组织的逻辑性与层次性

4.3　元素及其值的特征分析

元数据元素及其值的设计主要是为了实现数据集背景信息的完整描述、数据集权益的个性化声明以及数据集与相关资源的全面关联。科研人员为数据集创建的元数据将为数据重用与重现提供关键信息，其将决定数据何时可用、如何正确使用，同时也是数据引用的前提。对 6 个平台元数据元素及其值进行综合分析，其特点主要体现在 4 个方面，见表 5。

4.3.1　必备元素服务于数据引用

整体而言，6 个平台数据集发布所需的必备元素较少，SDB 则由于当前平台收录的数据集数量较少，为保证数据的检出，其所要求的必备元素相对较多。根据 2014 年发布的《数据引用原则联合声明》[18]，完整规范的数据引用格式中应当包括作者、年份、数据集标题、全球通用标识符、数据仓储名称、版本，而 Datacite 所鼓励使用的数据集引用格式中包括作者、出版年、标

表 5　元数据元素及其值的特征

数据仓储	数据引用（必备元素）	数据重用	数据民主控制	数据关联
DDR	较少	方法＋使用说明	CC0(强制性)	部分资源类型关联(论文、预印本、数据集、软件、补充信息):DOIs/URLs
Figshare	较少	在描述中陈述	1、多许可协议; 2、数据保护	多种资源关联:URLs/DOIs
HD	较少	在描述中陈述	1、CC0; 2、使用条款设置; 3、数据使用控制工具 Dataset Guestbook	1、多种资源关联:URLs; 2、特殊出版物关联:专门单元
Zenodo	较少	在描述中陈述	多许可协议	1、多种资源关联:URLs/DOIs 配合语义关联;2、特殊出版物关联:专门单元
MD	较少	描述＋重现步骤	多许可协议	多种资源关联:URLs/DOIs 配合语义关联
SDB	较多	数据集简介	1、CC0/CC-BY 4.0; 2、数据保护期	论文关联:DOIs/URLs

题、出版者、资源类型和标识符[15]。6 个平台要求用户所提交的必备元素集合基本包括数据集引用所必需的元素，未包括的元素通常在数据发布之时自动生成，如 DOI 和发布时间。

4.3.2　设定专门元素保证数据重用

为充分保证数据的重现和重用，平台均设置单独元素或者通过元素的组合要求用户在具体的值中提供数据收集的步骤与方法、所用仪器和设备等信息。按照陈述的具体性可分为简要陈述与详细陈述两大类：第一类通过在单一元素值下要求用户简述上述信息；第二类通过主元素值简要陈述配合副元素值完成上述信息的详细陈述，如 DDR 通过方法(Methods)与使用说明(Usage notes)这两个元素充分保证数据背景信息的完整呈现，MD 除描述(Description)这一必备元素外还通过专门设置重现步骤(Steps to reproduce)这一副元素要求用户详细陈述其数据收集过程。其中第一类陈述适用于平台内数据集体量较大时或者平台建设初期，第二类陈述主要适用于自然科学领域数据集。

4.3.3　元素组合机制实现数据民主控制

在 Wiley[1] 的调查中科研人员不愿意共享其科学数据的 2 个重要原因在于害怕数据滥用所产生的法律后果与缺乏对自己工作的认可机制。第一个原因产生的背景在于数据发布后科研人员失去对其数据集的控制，第二个原因产生的背景在于缺乏针对数据共享、数据发布的激励机制。针对上述问题，6 个平台主要采用 3 种方案实现数据灵活控制：①强制许可协议，如 DDR 所发布的数据集全部使用 CC0 许可协议，原因在于其数据集多与期刊论文关联且其主要服务于论文的同行评议；②多许可协议，其允许用户首先根据数据集类型灵活选择最适用的许可协议，如 MD 将协议分为纯数据、硬件与软件许可协议三大类，然后用户可在具体协议下自行决定数据集可被何种方式重用，包括署名、禁止商用、采用相同许可协议、禁止演绎以及通过组合方式进行重用限定；③许可协议与其他方式的组合使用，其允许用户在默认的少量许可协议之外，通过保护期首先设定数据保护的时间段或者公开的具体时间节点，如 SDB，而 Figshare 还可在上述基础上通过保护类型(Embargo type)允许用户选择对数据进行部分保护还是整体保护，若用户选择仅文件保护，则数据文件在保护期内处于私密状态，但其元数据记录将是可公开获取的，若用户选择整体内容保护，则数据集及其元数据记录在保护期间均是私密的，HD 则通过条款(Terms)这一复合元素允许用户从数据集层到文件层实现数据使用的具体限定，用户还可以通过内置工具 Dataset Guestbook 使得数据集被下载时数据集所有者能够收集下载者姓名、邮箱、机构和地理位置等关键信息。

4.3.4　多样化数据关联方式

各个平台均重视对数据集与其他关联资源之间关系的描述。按关联资源的类型可分为单一资源关联、部分资源关联与多种资源关联三大类，并主要通过 URLs 与 DOIs 两种方式实现数据集与多种类型的多个资源的精确关联。考虑到数据集与关联资源之间的关系类型同样具有多样性，包括引用、参考、汇编、衍生、部分、替代、连续和描述等关系，MD 和 Zenodo 尝试对关系进行语义化描述，并通过关联资源类型＋关联资源数字标志符＋关系三个子元素组配实现。此外，Zenodo 和 HD 设置专门单元为会议论文、学位论文、书籍或者报告的部分章节等特殊关联出版物提供适用的专门元数据单元帮助用户定位具体关联内容。

5　元数据创建服务的实现模式

目前，国内外科学数据元数据创建服务的形式主要有以下 6 种[9,19]：发布指导性文档、嵌入至科研过

程、自我提交表格、提供元数据文件模板、提供软件工具和智能解析元数据配置文件。按照人工参与程度，可分为完全人工创建、半手工创建、自动创建三类。科学数据仓储元数据创建服务的趋势则是根据平台自身的资源定位和用户需求，将以上几种服务形式融合，以实现科学数据元数据的自动化、智能化创建。

科学数据作为一种新兴学术资源逐渐被科研人员所接受。为解决当前科研人员认为的数据集元数据创建过程繁琐且工作量大、多数科研人员不知道或者不理解什么是科学数据的元数据且不知道该为自己的科学数据创建哪些元数据等问题，目前6个平台均通过自我提交表格并配合指导文档这一模式帮助用户完成科学数据的元数据创建。元数据创建方式以人工创建为主，半手工创建为辅，其中第一类固定选择性值为半手工创建，第二类外部参考值的创建方式介于半手工与完全人工创建之间，第三类用户著录值为完全人工创建。平台元数据创建服务实现模式既体现高效性、易用性与简洁性，同时也注重对科学数据元数据知识的普及。

5.1 自我提交表格

表格内容主要包括元数据元素名称、标识和解释三部分，根据元素解释覆盖元素的完整程度可分为元素完全解释型、元素部分解释型两类。平台以简洁的表格形式简化服务流程，借助清晰的符号标识为元素的重要程度分级并辅以具体的元素解释帮助用户无需系统学习专业的元数据知识即可快速进入并适应元数据创建环境。

6个平台对元素所进行的解释主要包括含义界定、规范著录示例、填写建议三大内容。与传统学术资源不同，为系统全面地描述科学数据的溯源信息以增加其重现性、可用性，数据集的描述通常会包括创建和转换历史元数据属性（Creation and transformation history metadata）以记录其如何创建以及对其所执行的后续转换、处理的信息[20]。部分解释型和完全解释型表格均对这一特殊属性所涉及的元素的含义加以简短、通俗的解释以增进用户对其的理解，从而引导用户正确、规范、完整地录入相关值，如表6所示：

表6　元素解释类型与特点

数据仓储	类型	具体特点
DDR	部分解释型	启发式设问：以具体问题引导用户思考和明确元素值所著录的内容
Zenodo	部分解释型	直接陈述：阐述每个具体元素及其值的意义和可资参考的内部与外部标准
Figshare	完全解释型	交互式解释：著录时弹出相应解释，提示主要包括元素含义、元素功能、著录内容、标准格式、操作步骤五部分
MD	完全解释型	交互式解释：著录时弹出相应解释，解释主要包括元素含义、元素功能、著录内容与要求三部分
SDB	完全解释型	交互式解释：著录时弹出相应解释，解释分为元素含义、元素功能、规范示例三部分
*HD	*完全解释型	*交互式解释，并对可能存在歧义、容易混淆的元素给出清晰定义与具体示例

具体而言，部分解释型平台采用启发式设问与直接陈述方式对相关元素进行解释与说明，主要体现了其对元素重要性的分级并充分考虑解释的易读与可读性；完全解释型平台则全部采用交互解释的方式对元素进行解释与说明，主要体现了解释的全面性并实现对用户困惑即时交互的解答。

R. NICOLAS[21]等分析了从DataCite收割来的7 440 415条元数据记录后发现用户对定义模糊、存在歧义的字段，会选择跳过不著录，因此有必要对存在歧义的字段清晰定义和解释，以提升元数据质量与完整性。HD则首次在交互式解释中对数据集生命周期中做出不同贡献的相关人员，如作者（Author）、生产者（Producer）、贡献者（Contributor）、分发者（Distributor）、存储者（Depositor）等元素，给出清晰定义，并对数据集中涉及的生成日期（Production date）、分发日期（Distribution date）、存储日期（Deposit date）、包含时间范围（Time period covered）、收集日期（Date of collection）等时间元素给出具体的概念定义与边界划分，最终实现元素意义的消歧。

5.2 指导性文档

表格内容为用户提供了关于元数据元素的必备基本知识，其内容多为操作层面的，平台提供单一或者多个指导性文档（见表7）满足用户进一步了解平台元数据信息的需求，以解决用户在元数据创建过程中遇到的实际问题。指导性文档一般设置在FAQ文档、帮助文档、数据管理与存档文档下，其内容涉及元数据的作用与意义、元数据中应包括的一般性内容、元数据所采用的标准与指导方案、补充说明每个元素的含义以及相关元素进行规范控制时所采用的术语系统。指导性文档旨在阐述什么是科学数据的元数据以及该为科学数据创建何种元数据，注重科学数据及科学数据元数据基础知识的普及以提升用户元数据创建能力。

表7　指导性文档内容

数据仓储	指导性文档
DDR[22]	FAQ：①应该在我的元数据中纳入什么信息？②在数据集提交之前我该如何准备数据文件？③数据集怎样才能变得可发现？
Figshare[23]	Help 文档中的 Tutorials 系列：①如何上载和发布你的数据集；②如何编辑和删除你的数据；③如何上载关联文件、保护和限制访问项目以及只上传元数据记录；④如何使用关键词组织你的文件＊(以上均附有操作视频)
HD[24]	用户指导：数据集与文件管理
Zenodo[25]	FAQ：①一般性问题；②技术与安全问题；③DOI 版本问题；④数据政策；⑤数据引用规范指导；⑥原则：FAIR 原则
MD[26]	数据集存档：① 数据摄取；② 数据管理
SDB[27]	Help 系列文档：①FAQ：如何填写你的数据集描述文件？②数据政策；③数据集发布流程

6　元数据创建服务特点

通过对6个综合型科学数据仓储的服务内容构成与实现模式进行系统剖析，尽管其提供的服务在微观层面上存在差异，但其整体上主要呈现以下特点：

6.1　充分吸收借鉴领域知识，沿袭传统并有所创新

综合型科学数据仓储在元数据元素及其值的设置上充分借鉴并吸收自然科学与社会科学领域元数据标准，立足于平台自身功能定位、元素规模与学科侧重点，在元数据的组织模式上将传统的自然科学与社会科学领域的模块化元数据组织方式进行改造并实现创新性表达。

6.2　以用户为中心，服务内容与模式上力求简洁

综合型科学数据仓储在服务内容方面摒弃自然科学与社会科学领域具体元数据标注方案的复杂性，其服务内容力求实现元数据内容的简洁与质量二者的平衡；在充分考虑不同层次知识背景用户体验的基础上，其服务模式力图实现元数据创建服务的简洁性、易用性、质量三者之间的平衡。通过内容与模式上的双重简化，实现科学数据的即时发布。

6.3　注重科学数据元数据知识普及，重视知识转化与迁移

在元数据创建实践中局部渗透科学数据元数据元素知识，在指导性文档中从整体上系统引入科学数据及其元数据背景知识，从而实现部分与整体知识的渗透与普及，并最终完成元数据知识与元数据创建能力的双向流动与互相促进。

6.4　注重对数据创建者的权益保护，充分体现数据民主

通过强制许可协议与多许可协议选择为数据的合理使用提供政策层面的支持。除利用基础许可协议实现数据的基本民主控制之外，配合用户自定义使用条款、保护期设定、内置开发工具，在基础之上进行改进与创新，进一步加强数据创建者权益保护以降低数据滥用风险。

6.5　重视数据集与其相关资源的关联组织，鼓励数据引用

V. TIMOTHY[28]等对发表于 *PLOS* 和 *BMC* 的50多万篇论文进行研究后发现，科研人员将自己的研究数据存储在数据仓储并与相关论文关联将使得论文的平均被引率提升约20%，《开放数据状态报告2019》[4]表明对研究论文的完整引用仍然是促使研究人员共享其科学数据的最强动力。各平台均提供专门的元数据元素或模块基本实现数据集与多种资源的关联，在基本关联之外，实现关联方式与关联表达上的创新。在数据集元数据创建完成之后，系统会自动生成多种数据引用格式，为数据集创造良好引用条件以提升数据集及相关学术成果影响力。

7　对我国图情机构数据仓储建设及元数据创建服务开展的启示

综合型科学数据仓储元数据创建服务从服务内容到实现模式上体现出上述诸多优势与特色，为我国图情机构科学数据仓储建设及科学数据元数据创建服务开展提供了先进的国内与国外经验，为此，可从以下3个方面学习借鉴并实现创新：

7.1　立足平台自身定位，服务内容力求简洁并凸显特色

我国图情机构在建设科学数据仓储时可该根据自身学科定位与主要服务对象，有针对性地选择、吸收、借鉴领域元数据标准，并在充分调研当下科研人员在数据集元数据标注中存在的困境与需求后，在力求简

洁的基础上形成具有自身特色的元数据标注与组织方案。数据仓储主要服务于自然科学领域科研人员期刊论文关联数据集存储与评审的,可参考 DDR 与 SDB 的元数据组织方案;数据仓储主要服务于社会科学领域科研人员数据集提交与存储的,可参考 HD 的将数据集元数据模块化与分层组织相结合的方案;追求极简模式并与出版商存在良好合作关系的多学科数据仓储,可参考 MD 与 Figshare 的元数据组织方案。

7.2　充分发挥图情机构信息素养培训优势,实现服务模式突破与创新

综合型科学数据仓储由于所存储数据集体量较大,数据集类型较多,其在实现模式上竭力做到服务的简洁性与易用性之间的平衡,但由于人员、经费以及管理资源的有限,综合型数据仓储仅通过指导性文档完成用户元数据创建的培训与指导,存在明显不足。我国图情机构可充分发挥机构在信息素养教育方面积累的丰富经验,开展科学数据素养培训与指导。为此,可在机构或者平台官网构建学习中心:①设置指导性文档板块对科学数据以及科学数据元数据知识进行系统梳理与总结,对用户在科学数据元数据创建过程中的历史问题进行归类、现存问题及时更新以及未来可能遇到的问题及时预判,并在问题后给出详细且可行的解决方案;②设置技能培训板块提升用户数据技能,定期发布线上、线下培训活动通知,并对视频、网络研讨会和书面指南等培训资源进行整理与归类,通过多样的培训形式与多种资源形式向用户展示如何检索、理解并正确使用数据集,最终提升用户元数据创建能力;③设置技术资源板块详细介绍当前科学数据分析、处理、搜集以及元数据创建中最常用的软件与工具,可通过发布工具使用指南与指向工具的链接使得用户了解并开始使用它们以提升元数据创建的效率与质量;④尝试设置科学数据管理板块,元数据方案作为数据管理计划 DMP(Data Management Plan,简称 DMP)的核心组成部分,自 2011 年美国国家科学基金会 NSF 提出 DMP 以来[29],越来越多的科研资助机构要求将数据管理计划作为基金与项目申请的必备材料,我国《科学数据管理办法》[30]中也明确要求各级科技计划管理部门建立验收科技计划的专项机制。未来我国图情机构可将科学数据的元数据创建服务、DMP 撰写服务、数据仓储的数据提交与存储服务相融合以充分发挥科学数据元数据创建服务在科研生命周期中的作用,可在数据收集前期通过元数据创建服务为 DMP 提供项目或基金申请所需的必要元数据,在数据收集阶段对元数据进行修正和补充,在项目结束后形成数据提交与存储所需的系统、完整的元数据。

7.3　利用图情机构工作人员专业知识优势,在元数据质量控制上进行创新

目前综合型科学数据仓储元数据创建主要以人工录入为主,以半自动化的元素值的选择录入为辅,面对海量的科学数据其尚未形成有效的元数据质量控制机制。2019 年 NIH 的数据科学战略办公室(NIH's Office of Data Science Strategy,简称 ODSS)与 Figshare 合作开展了一个为期一年的合作项目,其旨在确定生物医学领域研究人员如何使用 Figshare 来共享和重用 NIH 资助产生的科学数据,该项目发现数据集元数据经过专业数据馆员审查的数据集比未经审查的数据集下载量和浏览量高出 2.5 倍[31],这从侧面表明平台加强元数据质量审查的重要性。我国图情机构在开展科学数据的元数据创建服务时,可根据平台用户与数据集提交规模选择相应的元数据质量控制机制:规模较小的数据仓储可配置相应比例的数据馆员进行人工审核;用户规模与数据集提交规模较大的数据仓储可采用自动控制为主、人工审核为辅的质量控制机制,即开发相应的软件工具对数据集元数据形式与内容进行初步审核,并配置智能元数据评估系统完成对元数据质量的初步评估,数据馆员有针对性地对初步审核与评估中存在问题的数据集元数据进行二次审核并给出具体的补充说明与改进建议。

8　结语

本研究通过网络调研法并结合文献调研,从服务的内容构成与服务的实现模式两个方面对当前国际上影响力较大的 6 个综合型科学数据仓储所提供的元数据创建服务进行分析,总结出其元数据创建服务沿袭传统并有所创新、力求简洁并凸显自身特色、重视元数据知识普及与能力转化、充分保证数据民主、注重关联资源组织并鼓励数据引用五大特点,并为我国图情机构数据仓储建设及元数据创建服务开展给出具体参考建议与改进方法。

当前我国正在积极制定《数据论文出版元数据》国家标准[32],所选 6 个仓储作为数据出版的国际平台,其元数据创建服务体现了元数据标准典型应用场景,未来在制定标准时也可适当吸收借鉴综合型科学数据仓储在元数据元素选择、元素与值的设置方面的相关做法。

黄国彬，王涛. 综合型科学数据仓储元数据创建服务研究［J］. 图书情报工作,2021,65(21):131－140.

参考文献：

[1] FERGUSON L. How and why researchers share data (and why they don't)[EB/OL]. [2021－03－20]. https://doi.org/10.6084/m9.figshare.3468365.v1.

[2] TENOPIR C, ALLARD S, DOUGLASS K, et al. Data sharing by scientists: practices and perceptions[J]. PloS one, 2011, 6(6): e21101.

[3] TENOPIR C, DALTON E D, ALLARD S, et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide[J]. PloS one, 2015, 10(8): e0134826.

[4] Digital Science. The state of open data report 2019 [EB/OL]. [2021－04－27]. https://figshare.com/articles/report/The_State_of_Open_Data_Report_2019/9980783? backTo=/collections/State_of_Open_Data/4046897.

[5] 肖潇,吕俊生. E-science 环境下国外图书馆科学数据服务研究进展[J]. 图书情报工作,2012,56(17):53－58,114.

[6] 王翠萍,李佳璐. 国外高校图书馆科学数据服务现状与启示——以五所高校图书馆为例[J]. 图书馆工作与研究,2017(10):31－36.

[7] 钱鹏,郑建明. 高校科学数据组织与服务初探[J]. 情报理论与实践,2011,34(2):27－29.

[8] 邱春艳. 期刊文献与科学数据的关联服务研究[J]. 情报资料工作,2014(2):63－66.

[9] 黄鑫,邓仲华. 国外高校图书馆科学数据的元数据服务研究[J]. 图书与情报,2017(2):84－90.

[10] 胡芳. 国外典型科学数据仓储实施的元数据方案及启示[J]. 图书与情报,2015(1):117－121.

[11] Springer Nature. Research data policies[EB/OL]. [2021－01－21]. https: //www.springernature.com/gp/authors/research-data-policy/recommended-repositories/.

[12] Springer Nature. Generalist repositories[EB/OL]. [2020－12－09]. https://www.springernature.com/gp/authors/research-data-policy/repositories-general/12327166.

[13] 中国科学院计算机网络中心. ScienceDB 成为 Springer Nature 推荐通用存储库[EB/OL]. [2021－01－26]. https: //www.cas.cn/yx/202010/t20201010_4762415.shtml.

[14] DataCite Metadata Working Group. DataCite metadata schema documentation for the publication and citation of research data and other research outputs (Version 4.4). [EB/OL]. [2021－04－16]. https://doi.org/10.14454/3w3z-sa82.

[15] 杨波,胡立耘. 用于社会科学信息组织的元数据标准——DDI[J]. 现代图书情报技术,2005(8):7－11.

[16] UK Data Service. Quarterly labour force survey, April-June, 2021 [EB/OL]. [2021－09－03]. https://beta.ukdataservice.ac.uk/datacatalogue/studies/study? id=8826#! /details.

[17] LIANG D J, ZHUANG M H, HU C Y, et al. China's greenhouse gas emissions for cropping systems from 1978－2016[EB/OL]. [2021－09－04]. https://doi.org/10.1038/s41597-021-00960-5.

[18] Data Citation Synthesis Group. Joint declaration of data citation principles [EB/OL]. [2021－02－26]. https://www.force11.org/group/joint-declaration-data-citation-principles-final.

[19] 完颜邓邓. 国外科学数据仓储元数据实践调查及启示[J]. 新世纪图书馆,2016(5):81－84.

[20] SINGH G, BHARATHI S, CHERVENAK A, et al. A metadata catalog service for data intensive applications[EB/OL]. [2021－09－06]. https://dl.acm.org/doi/10.1145/1048935.1050184.

[21] ROBINSON N, JIMENEZ E, TORRES D. Analyzing data citation practices using the data citation index[J]. Journal of the Association for Information Science and Technology, 2016, 67(12): 2964－2975.

[22] Dryad. Frequently asked questions[EB/OL]. [2021－09－06]. https://datadryad.org/stash/faq.

[23] Figshare. Tutorials[EB/OL]. [2021－09－05]. https://help.figshare.com/section/tutorials.

[24] Harvard Dataverse. Dataset and file management[EB/OL]. [2021－09－05]. https://guides.dataverse.org/en/5.6/user/dataset-management.html.

[25] Zenodo. Frequently asked questions[EB/OL]. [2021－09－05]. https://help.zenodo.org/.

[26] Mendeley Data. How can we help you? [EB/OL]. [2021－09－06]. https://data.mendeley.com/faq.

[27] Science Data Bank. Frequently asked questions[EB/OL]. [2021－09－06]. https://www.scidb.cn/en/faq.

[28] VINES T H, ANDREW R L, BOCK D G, et al. Mandated data archiving greatly improves access to research data[EB/OL]. [2021－04－25]. https://doi.org/10.1096/fj.12-218164.

[29] NIH. Data management guidance for CISE proposals and awards [EB/OL]. [2021－02－02]. https://www.nsf.gov/cise/cise_dmp.jsp.

[30] 国务院办公厅. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. [2021－04－20]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[31] ANA V G. NIH figshare instance highlighted use cases[EB/OL]. [2021－09－07]. https://figshare.com/articles/online_resource/NIH_Figshare_Instance_Highlighted_Use_Cases/12816464.

[32] 中国科学院科学传播局. 国家标准《数据论文出版元数据》立项启动会召开[EB/OL]. [2021－04－09]. http://www.bsc.cas.cn/sjdt/202103/t20210324_4782140.html.

作者贡献说明：

黄国彬：提出研究思路与框架，论文最终版本修订；

王涛：选题确定，文献收集与整理，论文的撰写与修改。

Research on the Metadata Creation Service of the Generalist Research Data Repository

Huang Guobin Wang Tao
School of Government, Beijing Normal University, Beijing 100875

Abstract: [**Purpose/significance**] The lack of research data metadata knowledge and efficient easy-to-use metadata creation services hinder the sharing and reuse of research data among researchers. Due to the large amount and wide user orientation of the generalist research data repository, the metadata creation service provided by it has references for improving the above dilemmas. [**Method/process**] Taking six generalist research data repositories recommended by Springer Nature and Scientific Data as samples, this paper investigated and analyzed their metadata creation services from two aspects of the service content composition and the service implementation mode, then summarized their service characteristics and advanced experiences. [**Result/conclusion**] The metadata creation service provided by the generalist research data repository has five characteristics of following the tradition but has some innovations, striving for simplicity and highlighting its own characteristics, paying attention to the popularization of metadata knowledge and ability transformation, fully ensuring data democracy, paying attention to the organization of related resources and encouraging data reference. Its service model not only pays attention to the ease and usefulness of service but also taking account of the popularization of metadata knowledge, above which have important enlightenment and references for the research data repository construction and metadata creation service design for library and information community in China.

Keywords: research data metadata creation services the generalist repository